Utrecht University

Graduate School of Life Sciences

MSc Medical Imaging

2023 - 2024

Detection and Segmentation of Tissue Markers in Mammography: Studying their Impact on an AI Breast Cancer Detection System

Minor Research Project

Ana San Román Gaitero

Examination Committee:

Supervisor: Dr. Alejandro Rodriguez Ruiz
VP of Clinical Strategy, ScreenPoint Medical

Daily Supervisor: Daan Sperber
Clinical Data Scientist, ScreenPoint Medical

Examiner: Dr. Kenneth Gilhuijs
Associate Professor, UMC Utrecht

Detection and Segmentation of Tissue Markers in Mammography: Studying their Impact on an AI Breast Cancer Detection System

Ana San Román Gaitero
ScreenPoint Medical
a.sanromangaitero@students.uu.nl

***Abstract*—Breast cancer is the leading cause of cancer death among women. Accordingly, appropriate screening protocols for early diagnosis and treatment are crucial to fight against this disease. Artificial intelligence (AI) plays a significant role in screening mammograms and helping radiologists identify potential abnormalities. When a breast anomaly is detected, a biopsy is taken and a tissue marker is placed in the affected area, serving as a reference point for future treatments. To ensure reliability, it is crucial to comprehend how these tissue markers contribute to the process of breast cancer detection using AI systems. To this end, this study initially presents an image processing algorithm that facilitates the creation of an annotated dataset comprising images with tissue markers. This dataset is then used to develop a deep learning approach that is capable of discriminating mammograms with tissue markers, as well as of segmenting these objects. The study concludes with a methodology for analyzing the performance of an AI system in breast cancer screening. This framework does a comparison between the AI system's detected anomalous regions and the locations of tissue markers. The results highlight the strong performance of the deep learning model in both the segmentation and detection of mammograms with tissue markers. Moreover, the findings demonstrated that regardless of the presence of clips, AI systems can identify potential abnormalities with a reduced probability of marking tissue markers.**

***Index Terms*—Breast Cancer, Mammography, 2D X-ray, DBT, Tissue Markers, Surgical Clips**

I. Introduction

Breast cancer is the most diagnosed cancer and the leading cause of cancer-related death among women worldwide [1]. It is characterized by the presence of several lesions, including masses and microcalcifications, among others. Masses are large tumors that can be categorized as either malignant or benign, and microcalcifications are small calcium deposits that are an indicator of breast cancer or impending disease [2]. In the clinic, digital mammography is used as the standard breast cancer screening exam for the general population and has been proven to reduce breast cancer mortality [3]. This X-ray technique captures a two-dimensional (2D) image of the breast, enabling the detection of breast lesions, such as masses and microcalcifications, before they even become palpable. Identifying these lesions at an early stage allows for the improvement of survival rates through the early diagnosis and treatment of the disease [4]. Digital breast tomosynthesis (DBT) is a technique that is also used in the clinic to overcome some limitations of digital mammography, particularity with high breast density and overlapping tissue. Unlike 2D mammography, DBT acquires multiple projections from different angles along a predefined trajectory, which are then reconstructed into a series of high-resolution slices to obtain a three-dimensional (3D) image [5]. This method effectively addresses the limitations of 2D mammography and enhances the accuracy of lesion detection.

To assist radiologists improve the predictive performance of screening mammography, computer-aided diagnosis systems have been used for the past 20 years [6]. However, the exponential growth of artificial intelligence (AI) and the remarkable success in medical imaging have generated considerable interest in the development of deep learning algorithms to further enhance screening accuracy [6]. Over the years, several deep learning methods have been developed and shown to be efficient in solving breast cancer-related problems, including the detection of breast cancer lesions [7, 8], the identification of microcalcifications [9] and the segmentation of breast lesions [10, 11]. This shows that AI decision support systems can bring many benefits in facilitating radiologists' interpretation of mammograms for breast cancer detection. In particular in the context of cancer treatments, which can be highly effective when the cancer is diagnosed at an early stage.

In clinical practice, tissue markers, also referred to as surgical clips, are tiny objects commonly used to mark areas of interest during biopsy procedures [12]. The reason to place a clip within a breast lesion is mainly related to finding the area during surgery, after a biopsy, and for future treatments and follow-up [13]. The presence of these clips within mammograms presents a challenge for AI systems when assisting radiologists in detecting cancer, as they are placed in regions where suspicious masses or microcalcifications are present, which should be

detected by the AI algorithm. Clips are very bright and can be easily distinguished by the human eye, particularly by radiologists, who understand their significance when analyzing mammograms. However, AI systems lack the contextual understanding of clips and may struggle to identify masses accurately when clips are present. This could lead to errors when detecting breast cancer, such as marking clips as suspicious regions or influencing the decision-making process due to their presence. Therefore, it is important to study how AI breast cancer detection systems handle these specific cases in order to ensure reliability of the diagnostic result.

To date, no study has explored the influence of surgical clips on AI systems or the application of deep learning models to detect these tiny objects in mammograms. Nevertheless, research has been conducted on the detection of foreign objects in X-rays. Some studies have focused on the detection of pacemakers and defibrillators using models based on the MobileNet and DenseNet architectures [14], while others have investigated the detection of shoulder implants by integrating modified versions of ResNet-50 and DenseNet-201 [15]. Additionally, techniques for the segmentation or detection of small objects have also been developed, including the use of a multi-scale U-Net for the segmentation of abdominal small organs and lesions in computed tomography and magnetic resonance images [16]. A study also addressed the issue of incorrectly classified pixels by implementing a focal loss function to train a variation of a fully convolutional network for segmenting small stent graft markers in medical images [17].

The current study aims to investigate the impact of surgical clips on AI breast cancer detection systems, developing a deep learning algorithm capable of identifying and segmenting these small objects in 2D mammograms. An understanding of the interaction between AI systems and surgical clips could facilitate future advancements in AI-based breast cancer detection, potentially enhancing diagnostic accuracy in clinical settings.

II. Materials and methods

This section describes the research methodology and provides a comprehensive understanding of each step, illustrated in Figure 1. It begins with a description of the original data available and continues with the preprocessing techniques employed to obtain a clearer distinction between the two class labels required for the study. This is followed by the annotation process to obtain the clip segmentation masks and the data selection of the final dataset used for the subsequent task. Next, the architectural framework of the deep learning models employed in the detection of images with clips is explained. The last section summarizes the evaluation metrics used to analyze the robustness and performance of the experimental results.

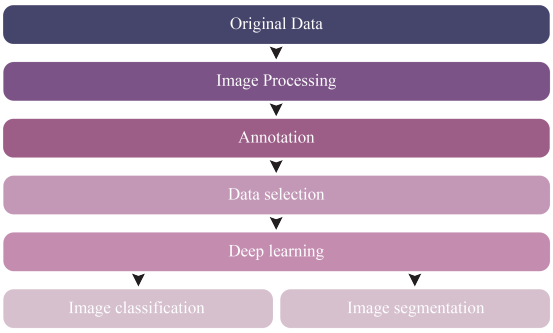

Fig. 1. Methodology workflow.

A Original Data

The provided dataset comprises a total of 3163 patients from two distinct sites, one in the Netherlands and the other in the USA. Each patient, also referred to as a case, could have undergone multiple examinations. Each exam includes one or more images, with each image showing a different perspective of the breast. The most common set consists of four images: the craniocaudal (CC) and mediolateral oblique (MLO) views for both the left and right breast. In this context, the study focused on individual images rather than entire examinations, given that not all image views in an exam contain tissue markers.

In particular, Site A comprises 16248 2D images, the majority of which were obtained by a Siemens device. In the exams with cancer, lesions were manually annotated based on the clinical reports under the supervision of a radiologist. Each annotation includes a lesion type (calcification group and/or soft tissue lesion), a cancer type (e.g. ductal carcinoma in situ (DCIS) or invasive lobular carcinoma (ILC)) and a region annotation with the location of the cancer. In contrast, site B comprised 37767 3D mammography images, also known as DBT, from a Hologic device. The study used synthetic images, which are 2D projections derived from the original DBT images.

Table I presents a summary of the dataset, including the number of patients, breast cancer outcomes, and images from each site. It is important to note that none of the images from the original dataset had been labeled with the presence or absence of tissue markers. Consequently, this

TABLE I
Original data

Dataset	Patients	N	B	M	Image Views
Total	3163	375	867	1921	54015
Site A	2038	0	617	1421	16248
Site B	1125	375	250	500	37767

Number of instances per variable. Patients with a normal (N), benign (B), and malignant (M) outcome.

dataset served as the starting point for the study.

B Image Processing

As previously stated, the provided dataset had no annotations on surgical clips, as well as any indication of the number of images that could contain one. Therefore, prior to training a deep learning model and in order to create a representative dataset for this task, it was necessary to label images containing tissue markers within the archive. Consequently, the first step was to develop and implement an automated image processing (IP) algorithm capable of discriminating images with clips from those without clips and generating their corresponding binary mask segmentations. The choice of image processing over other techniques was based on the high-intensity values of surgical clips, their distinct geometric shapes, and their contrast against surrounding tissues. In addition to the lack of training examples for any other AI algorithm. These characteristics suggested that image processing techniques could facilitate the annotation process. Later, a visual assessment and manual annotation were required, which resulted in the creation of a representative annotated dataset. This could then be used to improve the accuracy and efficiency of clip detection and segmentation through the application of deep learning models.

1. Algorithm

The IP algorithm comprised several phases, which are illustrated in Figure 2 and listed below. All functions were implemented with the Open CV library [18].

1. **Image normalization**. The initial step in the process was to normalize the data in order to ensure uniformity in the distribution of pixel values.
2. **Remove image background**. With the breast segmentations available, the next step involved removing the background from the breast, ensuring that no other image artifacts or image labels (such as the image view; MLO, or CC) could affect the results.
3. **Remove foreign objects other than clips**. One of the principal challenges encountered when applying enhancing techniques to highlight the bright areas was the presence of foreign objects with the same intensity values, such as pacemakers, wires, or stickers placed on the skin. To address this issue, it was necessary to apply more preprocessing techniques for each type of foreign object to obtain binary masks, which were then used to remove the objects from the breast image. This involved the application of several morphological operations, filters, and thresholds, with different kernel sizes and parameter values based on the foreign object under study (refer to Appendix A).
4. **Application of top hat filter**. Once the non-clip foreign objects had been removed from the image, the next step involved identifying those remaining structures with high-intensity values, which were presumed to correspond to the surgical clips. This was achieved by utilizing a top hat filter, with a kernel size of 3×3 and a rectangular structuring element, to enhance the brightest areas of the image.
5. **Application of morphological operations**. The next step implicated applying erosion and dilation techniques to isolate potential clip regions and remove other structures, such as small microcalcifications. The kernel size was set to 3×3 and an ellipse and a rectangle were set, respectively, as structuring elements.
6. **Application of mean shift filter**. A mean shift filter was applied to refine the pixel intensities within the resulting regions; pixel values within a clip could vary in a wide range of values making it harder to find the most optimal threshold. The mean shift filter was applied with a spatial radius of 25 and a range radius of 110.
7. **Thresholding**. A binary threshold was applied to the resulting image to obtain a binary mask of the corresponding regions, which in the best scenario should correspond with surgical clips. The threshold value was set to 129.
8. **Remove breast contour**. Due to the large number of instances with high-intensity pixels around the breast contour, a binary mask representative of this contour was obtained from the original image and aggregated with the previous binary mask to obtain the final segmentation mask. The breast contour mask was obtained by first finding the contour of the original breast mask and then dilating it with a kernel size of 8×8 for three iterations.
9. **Label each image**. The final stage assigned a label to each image. In order to distinguish between instances with and without clips, all regions within the final segmentation mask were identified and their area quantified. This area was used to determine whether the region in question corresponded to a clip or not. If the area of any of the regions in the image was below the specified threshold, the image was labeled as "clip". Conversely, if none of the contours were found to exceed the specified threshold, the image was labeled as "no clip". The range of area threshold was set between 81 and 800.

It should be noted that all parameters, including kernel sizes, types of morphological operations and filters, and applied thresholds, were selected based on the performance of the IP algorithm. These parameters were fine-tuned to optimize the segmentation of clip regions and ensure accurate detection of images with clips, using a manually selected set of fewer than 50 images. Specifically, to identify and select the optimal thresholds for the generation of the final binary mask and the distinction of clip regions from

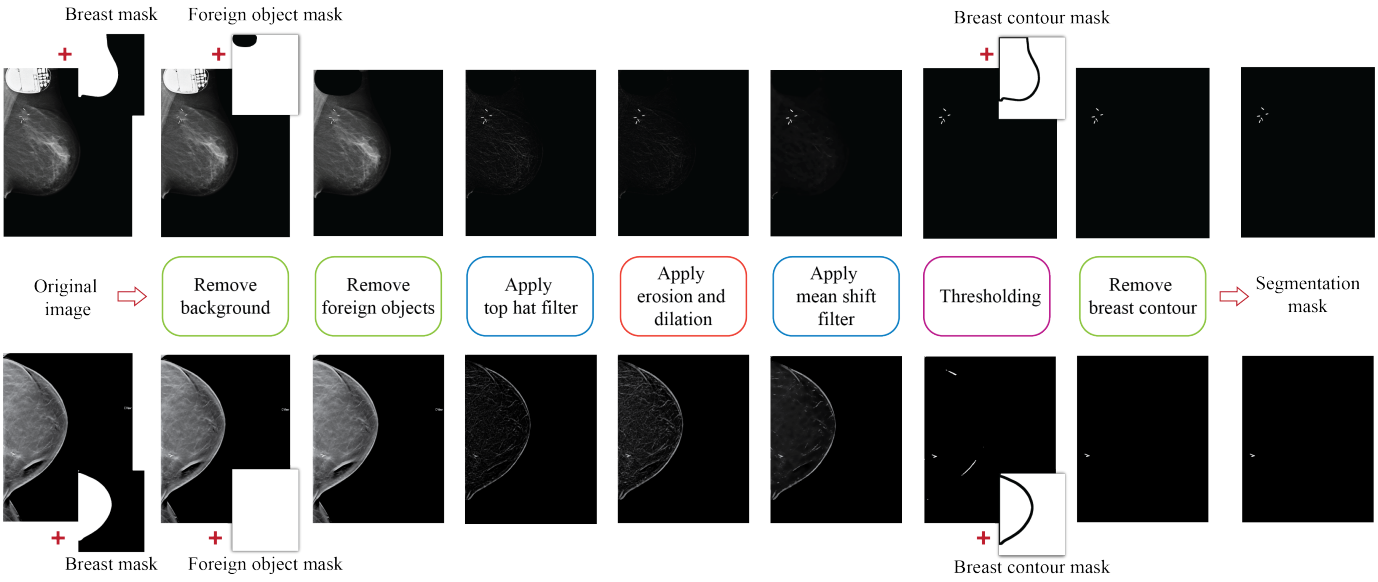

Fig. 2. Image processing framework. Two examples are shown with the resulting image above a description of the applied preprocessing step. Note that the first step, involving image normalization is not represented.

other regions, the following procedure was undertaken. The preprocessing algorithm was applied to all 50 images, and the resulting image from the mean shift filter output was obtained. Subsequently, a histogram of all these image outputs was generated, and the threshold was selected based on the optimal value that comprised the high-intensity pixel values while simultaneously removing other high-intensity values that were proximate to the clips. In contrast, the area threshold was determined by calculating the mean of all the areas of the regions that were found in the final masks and actually corresponded to clips.

2. *Application*

All images were processed by the IP algorithm in order to be classified into the two categories. The output images were paired with its binary segmentation mask. For those images without clips, the segmentation mask represented a zero image. Figure 3 illustrates different examples of the output segmentations obtained from the IP algorithm.

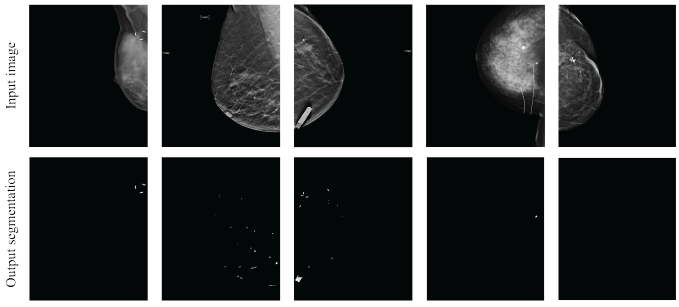

Fig. 3. Visual examples of the image processing algorithm outputs. Original images are seen on top, with their corresponding segmentation results on the bottom.

From left to right, the first image illustrates a successful segmentation, capturing all clips present in the image. The second image shows an image without clips, classified as such due to segmented microcalcifications and parts from the breast contour parts. The third image displays an example in which a pacemaker is present, which highlights the challenge of removing such objects to achieve optimal detection. However, this case was accurately classified as it contained clips. In contrast, the image on the right illustrates an example of a wire that has been successfully removed and a clip that has been accurately segmented. The last image shows a false negative (FN), an image that contains clips but has not been identified due to the absence of any segmentation.

C *Annotation*

The final dataset was obtained following a visual selection process, whereby any cases that had been misclassified were moved to the other class. Once the two classes had been distinguished, the corresponding binary segmentation masks were redefined by manual annotation using a Python graphical image annotation tool called LabelMe [19]. It is important to note that the annotation process was time-consuming, with each image containing between, approximately, one and twelve surgical clips that had to be manually segmented. Moreover, the majority of the images from the original dataset did not contain tissue markers, which resulted in a reduced number of images available for annotation and, consequently, for training the deep learning model.

D *Data selection*

The final dataset was a subset of the original set that included the images found with clips and a comparable

number of images without clips. This was done to ensure balanced data for training. Therefore, the data selection yielded a total of 3697 image views. This resulted in a final dataset comprising 2701 images from site A, of which 1083 contained clips, and 996 images from site B, of which 550 contained clips. Table II presents the clinical and demographic characteristics of the malignant cases, 1688 out of the 3697 patients from the study sample, used for the deep learning task.

TABLE II
Demographic and Clinical Characteristics

Characteristics	Sample (n=1688)
Cancer Region Characteristics	
Mass, n (%)	62.26%
Calcifications, n (%)	30.0%
Architectural Distortions, n (%)	3.55%
Asymmetries, n (%)	4.19%
Histological Cancer Types	
Invasive Non-Specific Type, n (%)	38.76%
Ductal Carcinoma In Situ, n (%)	11.63%
Invasive Lobular Carcinoma, n (%)	5.43%
Lesion Extent (maximum diameter)	
Lesion Diameter, median (IQR), mm	16 (11-25)
Demographic Characteristics	
Woman Age, median (IQR), years	61 (53-70)

IQR, interquartile range.

The dataset was divided into three sets for deep learning experimentation. A total of 50% of the images were allocated for training, 20% for validation, and 30% for testing (refer to Table III). A greater proportion of Site A was located into training because it contained easier instances where clips could be easily detected with respect to the background, which was observed in the IP outcomes. The test set included the majority of the images from Site B, which exhibited greater complexity and variability. The aim of prioritizing Site A for training and reserving a substantial portion of Site B for testing was to train the model on an easier clip detection task and to evaluate the model's performance under more realistic and challenging conditions found in clinical settings.

TABLE III
Data Split for Deep learning

Dataset	Site A	Site B
Train	1648	200
Validation	491	250
Test	562	546

Number of images for each dataset and site.

E Deep Learning

A deep learning algorithm was developed to identify images containing surgical clips across the entire data archive. Two different approaches were considered; a segmentation model and a classification model. All models were implemented with the Pytorch library using Python 3.8.10 [20].

1. Image Classification

Given that the primary goal is to simply identify which image contains clips and which does not, the first deep learning approach considered a classification model based on the ResNet50 architecture. The use of ResNet50 as a baseline allows for a consistent comparison of other techniques against a well-established standard model in the domain of classification tasks.

1.1 ResNet50

The ResNet50 architecture consists of 16 residual blocks, comprising 48 convolutional layers, followed by a max pooling layer and an average pooling layer [21]. The final layer is a softmax function that performs classification. The implementation of the ResNet50 model is conducted using the PyTorch library, in particular the package *torchvision.models*, using transfer learning. This approach employs pretrained weights to enhance performance and accelerate the convergence of the model. In order to adapt the ResNet50 for the binary classification between images with and without clips, a new linear layer with an output dimension of two was added to the final fully connected layer. As a result, the model is initialized using the pretrained weights from ImageNet [22], with all layers frozen except for those from the fourth block (layer 4) and the final fully connected layer. With this method, the final layer can be adapted to the tissue markers detection task while maintaining the valuable information obtained from the ImageNet feature representations.

1.2 Loss function

The ResNet50 classification employs a loss function that combines the standard cross-entropy loss with a weighted penalty for false positive (FP) predictions. This approach addressed an issue encountered in the initial runs where instances containing only microcalcifications were incorrectly labeled as clips. Therefore, by adding a 0.7 penalty to the cross-entropy loss, the model was able to reduce the FP predictions.

1.3 Experiments

A number of different learning rates, optimizers, and learning rate schedulers were studied in order to determine the optimal parameter configuration. The final ResNet50 model was trained for 50 epochs with a batch size of 32. The input images were resized to 224×224 pixels and

converted to a three-channel RGB format. Furthermore, as the model had been pretrained on the ImageNet dataset, the normalization process utilized the mean and standard deviation values associated with this dataset. In addition, a penalty weight of 0.7 was applied to the loss function, in combination with the AdamW optimizer and the ReduceLROnPlateau learning rate scheduler. The learning rate was set to 0.0005 and the weight decay to 0.01.

2. Image Segmentation

The next approach was undertaken for two reasons. Firstly, as previously stated, surgical clips are tiny objects located in areas where a biopsy has been taken, that is to say, areas where malignant or benign masses and microcalcifications may be present. These suspicious regions should be marked as high-risk areas for breast cancer. No study has been conducted to evaluate the performance of AI systems on images with clips; it is unclear whether they mark a clip because there is indeed a suspicious mass underneath, or because they have learned that a clip indicates a suspicious mass was present in the past, but there is no longer risk of cancer. Therefore, it is necessary to localize these clips in order to be compared with areas that the AI breast cancer detection systems have identified as potential lesions.

Secondly, as observed in the preprocessing step where the final dataset of images was obtained, clips were often mistaken with microcalcifications due to their high degree of similarity; small white spots that are randomly scattered or grouped like clips. This can make the classification task more challenging since a classification model lacks the morphological information required to accurately identify clips, specifically their appearance or location. For these two reasons, a segmentation approach was selected as the primary deep learning model, with a different twist on the usual approach. The study employed a U-Net architecture as a segmentation model, generating the corresponding segmentation masks and subsequently acting as a classifier according to the presence or absence of surgical clips.

2.1 U-Net

The U-Net architecture consists of an encoder (downsampling path), a bottleneck, and a decoder (upsampling path) [23]. Both the downsampling and upsampling paths apply two 3×3 convolutions with stride and padding values of 1, each followed by a ReLU activation. In the downsampling path, convolutional blocks are followed by a 2×2 max pooling operation with a stride of 2, reducing the spatial dimensions and progressively doubling the number of feature channels up to the bottleneck layer. It also uses skip connections to preserve the spatial information of the feature maps from the encoder to the decoder. The bottleneck layer includes two 3×3 convolutions followed by a ReLU activation, serving as a connection between the two streams. In the decoder path, transposed 2×2 convolutions upsample the feature maps, halving the number of feature channels. The upsampled feature maps are concatenated with the corresponding feature map from the encoder path, and retrieved from the skip connections, integrating high-level information with fine details. These concatenated maps undergo two 3×3 convolutions with ReLU activations. The final layer uses a 1×1 convolution to reduce the feature channels to the desired number for the final segmentation map, followed by a sigmoid activation for binary segmentation.

2.2 Loss function

The loss function employed for image segmentation is based on the one proposed by [24] and is shown in Equation 1. It is a combination of the Focal loss function and the Tversky index, which addresses the challenge of data imbalance, whereby tiny, small objects situated on a large background. Other techniques, such as Dice Similarity Coefficient (DSC) or weighted binary cross-entropy, encounter difficulties when dealing with this type of cases because they weigh FP and FN equally. In the context of highly imbalanced data and small regions of interest, such as surgical clips, it is better to assign greater weighting to FNs than to FPs in order to improve the recall rate. The Tversky similarity index addresses this issue by providing greater flexibility in balancing FPs and FNs through the introduction of two parameters, α and β. Additionally, the Focal loss introduces the parameter γ to prioritize hard examples, addressing the challenge of segmenting small objects due to their minimal contribution to the loss.

$$\text{Focal Tversky Loss} = \sum \left(1 - T_\alpha\right)^{\frac{1}{\gamma}} \tag{1}$$

In addition, a weight penalty is incorporated into the Focal Tversky Loss during training in order to address the challenge in the classification task of the segmentation model. The objective is to identify FP and FN by comparing the predicted segmentations with the ground truth labels (clip/no clip) at a pixel level. This enables the identification of areas where the model incorrectly predicts the presence of clips (positive) where none are present in reality (negative), and vice versa. The regions containing FP are multiplied by a penalty weight, denoted by θ, while those with FN are multiplied by the inverse of the penalty weight $(1 - \theta)$. The sum of these two penalties into a single metric results in the addition of the resulting value to the Focal Tversky Loss. The weighted penalty provides a detailed evaluation of the model's ability to distinguish between the two classes, namely the presence and absence of clips.

2.3 Experiments

For image segmentation the training process used images from both classes: clips and no clips, each paired with

their corresponding masks. This means that for images containing clips, the ground truth segmentation masks were binary, with clips represented by one-pixel values. On the other hand, for images without clips, the ground truth masks were formed by zero pixel values. Therefore, the segmentation model was forced to learn to generate a zero-valued segmentation when an image with no clips was input into the model.

Multiple experiments were carried out using a variety of criteria and approaches with respect to the classification performance. Some of these entailed the utilization of different image sizes and penalty weights for the loss function under evaluation. Moreover, a series of methods were conducted based on the model's predicted masks subsequent to a preliminary run with a reasonable parameter configuration. In this first run, the main challenge was the segmentation of regions containing microcalcifications, which, as negative instances, resulted in FP. Moreover, identifying clips located in bright areas or on top of microcalcifications and masses also led to a considerable amount of FN. To address this, further training approaches were implemented, including data augmentation techniques with Gaussian noise, rotation, and flipping. Furthermore, an additional penalty was introduced to the loss function when an image was predicted as a clip, but in reality, it contained only microcalcifications. This was achieved by adding a third label to the dataset, representing images containing only microcalcifications, to reduce FP. The final experiment involved the incorporation of images without clips into all three sets to evaluate whether the model's performance improved with a larger dataset.

After all these experiments and the identification of the optimal methodology and parameter configuration, the U-Net was trained using the 2D images and their corresponding masks as input, which were resized to a resolution of 512×427 and normalized between 0 and 1. The model was trained over 150 epochs using the AdamW optimizer with a learning rate of 0.0001, a batch size of 8, and a weight decay of 0.01. The Focal Tversky loss function was used to effectively deal with class imbalance, guiding the model optimization with $\alpha = 0.7$, $\beta = 0.3$, $\gamma = 0.8$, and $\theta = 0.6$. Early stopping was also implemented in order to prevent overfitting.

The output of the segmentation model comprised the predicted clip segmentations and the corresponding predicted labels, which were assigned to the predicted masks. This predicted label was calculated based on the predicted segmentation. If the sum of the mask pixel values equaled zero, it meant that no region was segmented and, consequently, no clip was identified in the image. Conversely, if the predicted segmentation included positive regions, it indicated that the model had identified clips within the image, and thus the "clips" label was assigned.

F Analysis and evaluation

In order to evaluate the performance of the different models, accuracy, precision, and recall were calculated from the ground truth and predicted labels in the classification task at the image level. The aforementioned calculations were performed for both the ResNet50 and U-Net model classifications. Furthermore, the same analysis was employed on the deep learning test set for the IP algorithm to enable a comparison with the deep learning models.

On the other hand, to assess the performance of U-Net image segmentation, the DSC was calculated at both the regional and image levels between the ground truth segmentations and the predicted masks. The image-level analysis evaluates the entire segmentation mask, considering the background, while the region-level analysis assesses the quality of individual clip segmentation within an image. The DSC is a metric that quantifies the overlap between two volume segmentations. It is defined as twice the intersection of the volumes divided by their union [25]. Therefore the DSC was employed to quantify the overlap between two clip segmentation masks. The approach initially determined the DSC across the entire test set at the image level, including both instances with and without clips. Later, it was calculated over only the set of images that contained clips, to provide a more accurate representation of the model's performance in clip segmentation. The region-level analysis was conducted by comparing each individual region identified in the prediction mask with each individual clip from the ground truth mask by calculating the DSC. Then, the best matches were identified by selecting the highest DSC from all possible combinations, as each clip had a single correspondence, resulting in one DSC value per clip.

III. AI in Breast cancer detection

The second part of the study used Transpara 2.1.0, an AI system developed by Screenpoint Medical BV in Nijmegen, the Netherlands. The aim was to evaluate the capacity of the AI system to identify suspicious masses and calcifications in mammograms with clips and to quantify their risk of cancer. Transpara analyzes each exam using a scoring system that ranges from 1 to 100, with 100 indicating the highest risk and 1 indicating the lowest risk of developing cancer [26]. Scores above 75 are classified as Elevated risk, with approximately 1 in 6 exams showing cancer. Scores between 74 and 43 are classified as Intermediate risk, with approximately 1 cancer in 250 exams. Lastly, scores below 43 are considered low risk with more than 4000 exams per cancer and which are not displayed to the user. As a result, Transpara only shows the user those findings that have been scored above 43 and provides a final exam score corresponding to the highest region score.

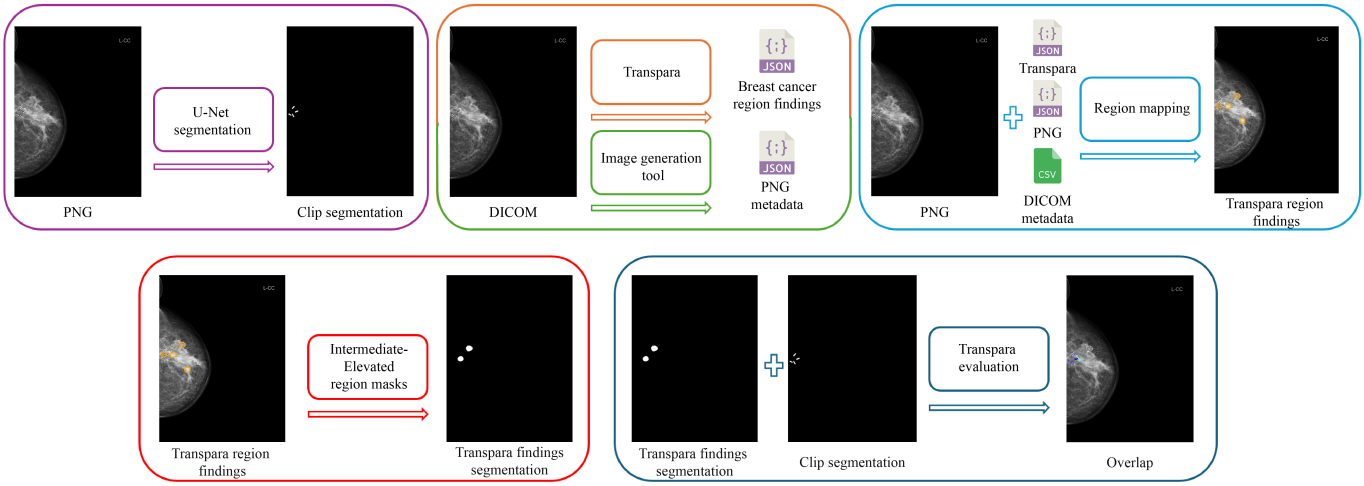

Fig. 4. Transpara's performance evaluation process. One example is shown with the resulting image next to a description of the applied step.

A *Performance Evaluation of Transpara*

In order to analyze how Transpara performs when clips are present in an exam, the study focused on individual findings categorized as Intermediate-Elevated (IE) risk of cancer. Note that these individual findings correspond to breast regions identified by Transpara. The following workflow was required due to the discrepancy between the deep learning model, which had been trained on PNGs, and Transpara's outputs, which are in DICOM format. Therefore, to align the model's output with Transpara's findings, it was necessary to translate the region findings from Transpara's coordinate system to the PNG coordinate system. Accordingly, the evaluation of Transpara's performance is illustrated in Figure 4 and involved the following steps:

1. **Clip segmentation using U-Net**: The best model weights of the U-Net obtained from the previous steps were used to process all the images in the original dataset. This step aimed to detect all images containing clips and to generate their respective clip segmentation masks.
2. **Transpara breast cancer detection**: Transpara processed the original dataset to identify anomalous regions and assign corresponding risk scores in 2D and DBT DICOM exams. This step produced JSON files with the region-finding results, including contour pixel coordinates and their region scores. The spacing, translation, and flipping information of the original DICOM images were obtained from the ScreenPoint database.
3. **Image generation**: An in-house tool was used to generate the PNG images from the DICOM images in the original dataset. This step generated the corresponding JSON files containing the spacing, translation, and flipping information of the PNG image. This was required to convert the image from DICOM space to PNG.
4. **Region mapping**: An interpolation process was performed once the Transpara findings and the generated PNG JSON files were obtained. This entailed the processing of JSON files from both algorithms, the extraction of region details, and the translation of coordinates based on spacing, translation and flip information to accurately map Transpara findings from DICOM onto the PNG images.
5. **Selection of high-risk regions (IE risk)**: Binary segmentation masks were created for those findings that have region risk scores above 43, which correspond to those regions with an IE risk of breast cancer, as previously explained.
6. **Transpara evaluation**: The final step involved comparing the images that contained clips and IE regions found by Transpara. To analyze the degree of overlap between both areas, a region-level DSC approach analogous to the one employed by the U-Net was performed. This entailed computing a DSC for each combination of clip-finding in both masks. Given that the goal was to determine the likelihood of Transpara marking a clip as a lesion, only the match pair with the highest DSC was taken into account. If no match was found, it could be inferred that Transpara had not marked a clip as a lesion. Conversely, if a matching pair was identified (any DSC between both regions), it could be concluded that Transpara had marked a clip itself or a nearby lesion.

IV. Results

In this section, the results of the experiments are presented. It is important to note that the IP algorithm was not intended to be a final segmentation or detection

method but is included in Results Section A and Section B to highlight and discuss the differences between traditional algorithms and AI-based approaches.

A Image Classification

The results of all image classification methods are presented in Table IV. Firstly, the experimental results demonstrate that transfer learning enhances ResNet50's accuracy performance, reaching 0.85. As expected, the U-Net outperforms the other methods in terms of accuracy, recall and precision, achieving 0.95, 0.93 and 0.97 respectively. Conversely, the IP algorithm achieved the lowest accuracy of all, reaching 0.78, but a higher recall rate in comparison to the ResNet-50 model. This is because the IP method served as the baseline method to create the final dataset, thus, it was expected to have a high performance.

TABLE IV
Image Classification Performance

	Evaluation Metrics (95% CI)		
Model	**Accuracy**	**Recall**	**Precision**
U-Net	**0.948** (0.93-0.96)	**0.932** (0.91-0.95)	**0.947** (0.93-0.97)
ResNet50	0.847 (0.83-0.87)	0.787 (0.75-0.82)	0.852 (0.82-0.89)
IP	0.782 (0.75-0.80)	0.906 (0.88-0.93)	0.694 (0.65-0.73)

Evaluation metrics with a 95% confidence interval (CI). Image processing algorithm (IP).

Moreover, Table V presents the accuracy, recall and precision on both data sites separately. The U-Net demonstrated a strong predictive performance reaching 0.98 accuracy in Site A, and a high sensitivity and precision in the detection of clips, with a score of 0.99 and 0.94, respectively. Furthermore, the U-Net also reaches the highest accuracy and precision in Site B, with 0.92 and 0.95, respectively. Conversely, ResNet50 demonstrates robust accuracy for tissue marker classification in Site A, with a value of 0.94. However, its performance declines when predicting instances from Site B, with an accuracy of 0.75. Lastly, as expected, the IP algorithm demonstrated the lowest performance in comparison with the other two deep learning models, although the sensitivity towards Site B was the highest of all three methods, with a value of 0.92.

B Image Segmentation

The performance of the U-Net model in image segmentation was initially evaluated by the DSC at the image level for the entire test set, which included images with and without clips. The U-Net achieved an overall DSC of 0.84 in the test set. With regard to Site A, the model achieved a DSC of 0.90, whereas for Site B, the DSC was 0.78. Moreover, to assess the efficacy of the U-Net in segmenting the clips, the evaluation was also conducted exclusively on images with clips at an image and region level. The results can be seen in Table VI, which displays the median DSC and interquartile range for the entire dataset comprising clips in both methods. The median DSC of the U-Net at both the image and region levels was 0.85 and 0.88, respectively. In contrast, the IP algorithm exhibited substantially lower DSC, with a median of 0.55 and 0.66 at the image and region levels, respectively.

TABLE VI
Image segmentation Performance

Methods	**Image-Level**	**Region-Level**
U-Net	**0.85** (0.75-0.90)	**0.88** (0.80-0.92)
IP	0.55 (0.31-0.69)	0.61 (0.00-0.75)

Dice Similarity Coefficient (DSC) Median and Interquartile Range (IQR) on both image and region levels for the entire dataset containing clips. Image processing algorithm (IP).

TABLE V
Image Classification Performance per Datasets

		Evaluation Metrics (95% CI)		
Dataset	**Model**	**Accuracy**	**Recall**	**Precision**
Site A	U-Net	**0.975** (0.962-0.988)	**0.989** (0.981-0.998)	**0.938** (0.918-0.958)
	ResNet50	0.933 (0.911-0.955)	0.900 (0.874-0.927)	0.933 (0.911-0.955)
	IP	0.895 (0.869-0.920)	0.888 (0.842-0.934)	0.812 (0.757-0.866)
Site B	U-Net	**0.919** (0.897-0.942)	0.897 (0.872-0.923)	**0.954** (0.937-0.972)
	ResNet50	0.788 (0.737-0.839)	0.800 (0.750-0.850)	0.708 (0.652-0.764)
	IP	0.669 (0.630-0.709)	**0.917** (0.885-0.948)	0.640 (0.594-0.685)

Evaluation metrics and 95% confidence intervals (CI) achieved with each Site in the classification tasks. Image processing algorithm (IP).

The analysis was also conducted separately for each site, as illustrated in Figure 5, only in images containing clips. The following figure presents violin plots of the DSC for the U-NET and the IP algorithm, evaluated on both image and region levels. Detailed statistical information on the DSC distributions is provided in Appendix B. The U-NET demonstrated superior and more reliable segmentation at the image level, as evidenced by a more concentrated distribution towards higher DSC compared to the IP method. The broader DSC distribution for the IP algorithm indicates inconsistency in achieving accurate segmentation across different images and clips. Moreover, Site A had a greater number of higher DSC values compared to Site B in both methods. It is noteworthy that at a region level, two distinct clusters of DSC scores can be observed, one towards higher DSC values and one towards zero. In particular, the U-Net has the majority of the values concentrated in the highest peak in both sites, indicating that it is successful in segmenting all the clips present in the image. In contrast, the IP has a higher proportion of values around zero in both sites, suggesting that it does not capture all the clips.

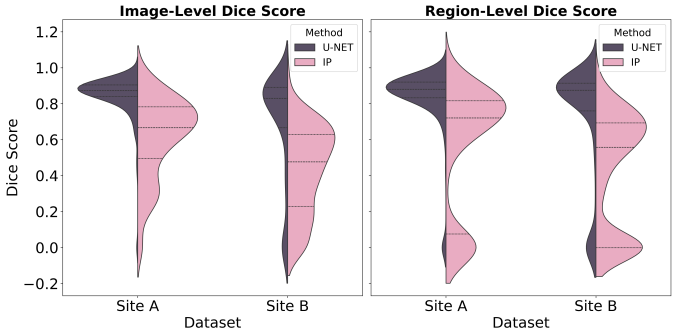

Fig. 5. Comparison of DSC for image and region level analysis across both sites and segmentation methods. IP, Image Processing

C *Transpara's performance evaluation*

After processing the original data, the study detected a total of 2696 exams (6160 image views) containing clips using the U-Net segmentation model. However, due to some limitations of the in-house image generation tool, the number of exams with clips and Transpara findings was reduced to 1154 exams (1956 image views). Thus, the final study population consisted of 1154 exams. After the region mapping process and selection of high-risk regions, 753 exams were found to be in the IE category, indicating that 401 exams, representing 34.7% of the total, exhibited no evidence of lesion detection. Finally, in Transpara's evaluation step, the number of exams that exhibited an overlap between at least one of the findings and a clip was 182 (226 image views), seen in Figure 6. This suggests that in 571 exams, or 50% of the time, a clip does not overlap with an IE risk finding.

As illustrated in Figure 7, the scenarios were diverse. Starting from the top left, the first image shows the scenario

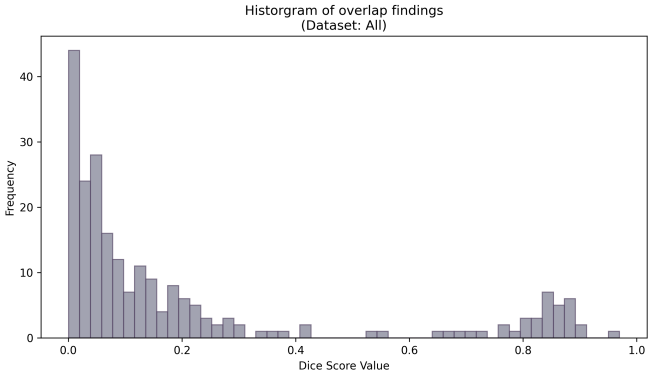

Fig. 6. Histogram of the DSC overlap between the clip segmentation mask and Transpara region mask

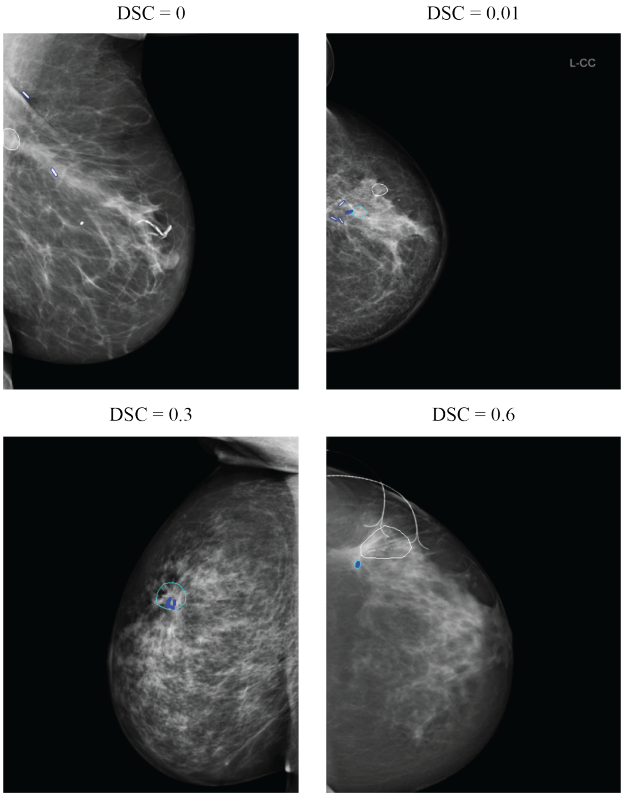

Fig. 7. Example of different image views and different DSC after the evaluation of Transpara. The outcome represented by a light blue Transpara finding and a dark blue filled clip is an overlap. In contrast, a non-overlap outcome is illustrated as a white finding and an unfilled blue clip.

in which a clip and a finding do not have any overlap. Furthermore, in instances where the DSC value fell within the range of 0 to 0.01, as seen in the top right image, the clip was found to be in contact with the region delineated by Transpara, but not within it. If the DSC value was in the range of 0 to 0.6, the clip was either partially or completely within a region marked by Transpara. It is noteworthy that in this case, surrounding tissue was also present, as

illustrated in the bottom left image. Nevertheless, when the DSC was equal to or greater than 0.6, Transpara marked the entire clip as a finding, indicating a false positive in the detection of cancer (see the bottom right image in Figure 7). Consequently, in our study population of 1154 exams, the results show that Transpara incorrectly identifies a clip as an IE finding in 2.8% of cases (32 exams). Nevertheless, this also indicates that while there is a 2.8% chance of FP due to clips, Transpara has a higher accuracy of 13% (150 exams) in identifying and marking abnormal tissue with a high risk of breast cancer, regardless of the presence of a clip.

V. Discussion

Although the IP algorithm was not originally designed for the purpose of detecting and segmenting clips in 2D mammograms, it proved to be a valuable tool for the creation of an adequate dataset. The algorithm demonstrated its potential as a first step to identify images containing clips and to facilitate manual selection and annotation of the dataset. Nonetheless, some challenges were found during the process. The inconsistency in pixel intensity values across different images, as well as between clips within the same image, led to difficulties in optimizing threshold values and generalization processes. Foreign objects other than clips, such as pacemakers and wires, also posed challenges due to similar intensity values, as seen in Figure 3. Even though their larger areas facilitated partial removal in the initial stages, some residuals were still present in the final segmentation masks. However, it is noteworthy that, for the purpose of this algorithm, prioritizing FP over FN was acceptable. This approach ensured that images containing clips were not missed, even if it meant additional work to remove incorrectly identified microcalcifications or other artifacts during the manual annotation. Lastly, if this algorithm were to be improved for a more autonomous task of detecting and segmenting clips across images, a more larger and diverse dataset would be required to make a more extensive optimization experiment. It is, however, important to note that, as an IP algorithm constrained by predefined operations and filters, its ability to improve performance may be constrained by inherent limitations.

It is noteworthy that microcalcifications were the most challenging tissue type within the breast to differentiate from surgical clips due to their high degree of similarity. This was encountered in the IP algorithm, seen in Figure 3, as well as in the deep learning models. This was expected, given that these calcium deposits have similar intensity values and occasionally rounded shapes, as some of the tissue markers observed in mammograms. To address this issue, it was necessary to incorporate an additional FP penalty into the deep learning models. Future work could include additional microcalcification annotations to improve clip segmentation and differentiation between these two objects.

In regard to the deep learning classification models, as expected, the U-Net model demonstrated superior performance in detecting images with clips. This may be attributed to the fact that, although a segmentation model is not designed for classification purposes, the additional spatial information provides extra information regarding the appearance of clips, facilitating the differentiation from other similar regions, such as microcalcifications. Additionally, ResNet50 reached a high accuracy in detecting clips, particularly in those cases from Site A. However, it consistently performed below U-Net, which may be attributed to the deep network nature of ResNet50 and the fact that the dataset was not sufficiently large to enable fine-tuning of the model. It is notable that ResNet50 also encounters greater difficulty in identifying clips from Site B, which suggests that it is more challenging to predict complex examples for the model. On the other hand, U-Net demonstrated its capability to learn clip features mostly from Site A and to identify with high-certainty clips within Site B, therefore, reaching a strong level of generalization. This shows that U-Net is capable of detecting clips in more complex and diverse clinical contexts, so if other sites from more diverse backgrounds were to be predicted by this model, the chances of achieving strong performance would be higher.

The same observations can be derived from Figure 5 regarding the segmentation results and in comparison with the IP algorithm. At the region level, the U-Net achieves DSC values between 0.8 and 1 for the majority of clips in both sites. In addition, the broader distribution of the IP algorithm in the image-level analysis suggests that its segmentation performance is inconsistent, with significant variability likely due to partial clip segmentations or errors in differentiating clips from similar objects, which was seen in Figure 3. Moreover, the distribution on the region level demonstrates that clips are rarely fully segmented, with the majority of instances exhibiting incomplete segmentation, as evidenced by the highest peak of the distribution, which has DSC values between 0.5 and 0.8. This may be attributed to the intensity discrepancy within regions previously discussed, which makes it challenging to determine an optimal threshold. Overall, the results demonstrate that the U-Net is effective under more realistic and challenging scenarios that can be found in clinical settings. It accurately detects and segments clips with a superior performance compared to the ResNet50 and also the IP algorithm.

The study evaluated Transpara's performance in breast cancer detection on 1154 exams containing clips. Transpara exhibited a false positive rate of approximately 3%, yet identified anomalous regions with a high risk of breast cancer in 13% of cases, regardless of the clips. It is evident that the higher accuracy in detecting potential cancers in situations where Transpara might have more challenges outweighs the risk of marking 3% of the clips. Additionally, the presence of 50% of exams with no overlap

between clips and Transpara findings and the 35% of the cases with no abnormalities detected further demonstrates Transpara's robust performance in scenarios where clips are present. In the context of cancer detection, and given that Transpara assists radiologists, this trade-off may be considered acceptable. Regions that have been misclassified, such as clips, can be reviewed by radiologists, meanwhile, the system ensures that masses are not being overlooked. Prioritizing the identification of potential malignancies over the risk of FP is crucial in the diagnosis of breast cancer. To further enhance Transpara's performance, incorporating the U-Net pipeline to detect the presence of clips and compare the clip masks with Transpara's finding masks could reduce the FP rate. This improvement could enhance the reliability of Transpara's algorithm in distinguishing true positives from false positives involving clips.

Moreover, it should be noted that the majority of images from Site B were not used to evaluate the performance of Transpara due to limitations in the intermediate steps. Most of these images were DBT, an image format that the in-house image generation tool was not able to process. As a result, out of nearly 5000 images that were found to contain clips within this site, only 500, which were obtained from 2D mammograms, were available for the analysis. Addressing this issue could potentially increase the number of images processed, thereby providing a larger dataset for the evaluation. Despite this, it is unlikely that enlarging the dataset will alter the results of Transpara, as the dataset used was already considered to have high quality and variability, ensuring a reliable and robust evaluation.

Some limitations were encountered in this study. The annotations for the tissue markers were not available, requiring manual annotation to redefine the masks obtained from the IP algorithm. This introduces user variability, given that the annotations were not checked by a radiologist. Future work could consider involving experts to ensure consistent annotations. On the other hand, exploring nnU-Net or other models such as transformers, could potentially improve performance and robustness when dealing with dual tasks, such as simultaneous segmentation and classification. Lastly, this study only considered 2D images, thus, incorporating 3D images could enhance overall performance. Instead of working with synthetic DBT images and digital mammography, the incorporation of a third dimension could enhance the understanding of how surgical clips are and maybe help in differentiate them from microcalcifications, for instance.

In the context of a woman who has undergone biopsy and treatment, information on whether a region with a clip was previously diagnosed as malignant and has now been diagnosed differently by an AI system in the following examinations should be further researched. This could provide information about post-treatment changes or potential recurrence of malignancy, since radiologists may have difficulty interpreting follow-up mammography due to these changes [27]. In cases where treatment is successful, the affected tissue should become benign and no detection of malignancy should be reported. With the information on surgical clips that was obtained in this study, it could be valuable to study how different treatments vary the effects on the surrounding tissue and see what treatments are more effective in reducing malignancy. Therefore, longitudinal studies in regions with clips can provide information to radiologists to redefine treatment or personalize therapy. However, to make this possible more information would be needed, such as the pathology results from a biopsy. Furthermore, an important area of future research is investigating how cancer predictions are obtained under regions containing surgical clips. This involves checking whether these regions are being marked as malignant or benign because of the presence of the clip. Explainable AI techniques, such as saliency maps of the decision-making process, can highlight the regions contributing to the model's prediction [28]. If the clips are significant contributors to the malignant or benign outcome, it may indicate that the clip is triggering the model's decision and affecting the predictions. Therefore, the incorporation of surgical clip information into the radiologists' report, and the investigation of all these suggestions, could potentially enhance the contribution to better clinical outcomes.

VI. Conclusion

The current study explores the impact of tissue markers on the performance of AI systems in breast cancer detection by introducing a deep learning algorithm that can not only identify but also segment these tiny objects in mammography. The study shows that these models are suitable for the detection of tissue markers, achieving strong performance in the segmentation of all clips within a 2D mammogram, even in the presence of more complex and variable data. Moreover, despite the presence of clips, the AI system under evaluation is able to identify areas of concern that may require further attention and exhibits a low incidence of marking clips as potential abnormalities. Thus, indicating that it is robust to tissue markers when assisting radiologists in breast cancer detection. Nevertheless, the study of tissue markers in breast cancer detection and AI systems remains a challenge. If these systems were to become fully autonomous, further improvements should be examined to enhance their performance with surgical clips. Future work should focus on exploring the impact of tissue markers on the model's decision-making process, ensuring predictions are not influenced by these objects. Such advancements could facilitate a deeper understanding and management of AI in breast cancer detection.

References

[1] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers

in 185 countries". In: *CA: a cancer journal for clinicians* 71.3 (2021), pp. 209–249.

[2] José Daniel López-Cabrera, Luis Alberto López Rodríguez, and Marlén Pérez-Díaz. "Classification of breast cancer from digital mammography using deep learning". In: *Inteligencia Artificial* 23.65 (2020), pp. 56–66.

[3] László Tabár, Bedrich Vitak, Tony Hsiu-Hsi Chen, Amy Ming-Fang Yen, Anders Cohen, Tibor Tot, Sherry Yueh-Hsia Chiu, Sam Li-Sheng Chen, Jean Ching-Yuan Fann, Johan Rosell, et al. "Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades". In: *Radiology* 260.3 (2011), pp. 658–663.

[4] Kwan Hoong Ng and Malai Muttarak. "Advances in Mammography Have Improved Early Detection of Breast Cancer". In: 2003. URL: https://api.semanticscholar.org/CorpusID:35022368.

[5] Luca Nicosia, Giulia Gnocchi, Ilaria Gorini, Massimo Venturini, Federico Fontana, Filippo Pesapane, Ida Abiuso, Anna Carla Bozzini, Maria Pizzamiglio, Antuono Latronico, et al. "History of mammography: analysis of breast imaging diagnostic achievements over the last century". In: *Healthcare*. Vol. 11. 11. MDPI. 2023, p. 1596.

[6] Li Shen, Laurie R Margolies, Joseph H Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh. "Deep learning to improve breast cancer detection on screening mammography". In: *Scientific reports* 9.1 (2019), p. 12495.

[7] Nusrat Mohi ud din, Rayees Ahmad Dar, Muzafar Rasool, and Assif Assad. "Breast cancer detection using deep learning: Datasets, methods, and challenges ahead". In: *Computers in Biology and Medicine* 149 (2022), p. 106073. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2022.106073. URL: https://www.sciencedirect.com/science/article/pii/S0010482522007818.

[8] Ghulam Murtaza, Liyana Shuib, Ainuddin Wahid Abdul Wahab, Ghulam Mujtaba, Henry Friday Nweke, Mohammed Ali Al-garadi, Fariha Zulfiqar, Ghulam Raza, and Nor Aniza Azmi. "Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges". In: *Artificial Intelligence Review* 53 (2020), pp. 1655–1720.

[9] Daesung Kang, Hye Mi Gweon, Na Lae Eun, Ji Hyun Youk, Jeong-Ah Kim, and Eun Ju Son. "A convolutional deep learning model for improving mammographic breast-microcalcification diagnosis". In: *Scientific reports* 11.1 (2021), p. 23925.

[10] Wessam M Salama and Moustafa H Aly. "Deep learning in mammography images segmentation and classification: Automated CNN approach". In: *Alexandria Engineering Journal* 60.5 (2021), pp. 4701–4709.

[11] Michiel Kallenberg, Kersten Petersen, Mads Nielsen, Andrew Y Ng, Pengfei Diao, Christian Igel, Celine M Vachon, Katharina Holland, Rikke Rass Winkel, Nico Karssemeijer, et al. "Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring". In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1322–1331.

[12] Ami D Shah, Anita K Mehta, Nishi Talati, Rachel Brem, and Laurie R Margolies. "Breast tissue markers: Why? What's out there? How do I choose?" In: *Clinical imaging* 52 (2018), pp. 123–136.

[13] Inyoung Youn, Seon Hyeong Choi, Shin Ho Kook, Yoon Jung Choi, Chan Heun Park, Yong Lai Park, and Dong Hoon Kim. "Ultrasonography-Guided Surgical Clip Placement for Tumor Localization in Patients Undergoing Neoadjuvant Chemotherapy for Breast Cancer". In: *Journal of Breast Cancer* 18 (2015), pp. 44–49. URL: https://api.semanticscholar.org/CorpusID:14893040.

[14] Ue-Hwan Kim, Moon Young Kim, Eun-Ah Park, Whal Lee, Woo-Hyun Lim, Hack-Lyoung Kim, Sohee Oh, and Kwang Nam Jin. "Deep learning-based algorithm for the detection and characterization of MRI safety of cardiac implantable electronic devices on chest radiographs". In: *Korean Journal of Radiology* 22.11 (2021), p. 1918.

[15] Haseeb Sultan, Muhammad Owais, Chanhum Park, Tahir Mahmood, Adnan Haider, and Kang Ryoung Park. "Artificial Intelligence-Based Recognition of Different Types of Shoulder Implants in X-ray Scans Based on Dense Residual Ensemble-Network for Personalized Medicine". In: *Journal of Personalized Medicine* 11.6 (2021). ISSN: 2075-4426. DOI: 10.3390/jpm11060482. URL: https://www.mdpi.com/2075-4426/11/6/482.

[16] Lei Wang, Bo Wang, and Zhenghua Xu. "Tumor Segmentation Based on Deeply Supervised Multi-Scale U-Net". In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (2019), pp. 746–749. URL: https://api.semanticscholar.org/CorpusID:211056299.

[17] Xiao-Yun Zhou, Mali Shen, Celia V. Riga, Guang-Zhong Yang, and Su-Lin Lee. "Focal FCN: Towards Small Object Segmentation with Limited Training Data". In: *ArXiv* abs/1711.01506 (2017). URL: https://api.semanticscholar.org/CorpusID:195347076.

[18] Itseez. *Open Source Computer Vision Library*. https://github.com/itseez/opencv. 2015.

[19] Kentaro Wada. *labelme: Image Polygonal Annotation with Python*. https://github.com/wkentaro/labelme. 2018.

[20] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer. 2015, pp. 234–241.

[24] Nabila Abraham and Naimul Mefraz Khan. "A novel focal tversky loss function with improved attention u-net for lesion segmentation". In: *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. IEEE. 2019, pp. 683–687.

[25] Abdel Aziz Taha and Allan Hanbury. "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool". In: *BMC medical imaging* 15 (2015), pp. 1–28.

[26] Alejandro Rodríguez-Ruiz and Nico Karssemeijer. "Artificial Intelligence to Help Radiologists in the Early Detection of Breast Cancer with Mammography and Breast Tomosynthesis". In: *Breast Care. Because We Care.* (2020), p. 15.

[27] Misugi Urano, Hiroko Nishikawa, Taeko Goto, Norio Shiraki, Masayuki Matsuo, Fatmaelzahraa Abdelfattah Denewar, Naoto Kondo, Tatsuya Toyama, and Yuta Shibamoto. "Digital Mammographic Features of Breast Cancer Recurrences and Benign Lesions Mimicking Malignancy Following Breast-Conserving Surgery and Radiation Therapy." In: *The Kurume medical journal* (2018). URL: https://api.semanticscholar.org/CorpusID:208018504.

[28] Said Pertuz, David Ortega, Érika Suarez, William Cancino, Gerson Africano, Irina Rinta-Kiikka, Otso Arponen, Sara Paris, and Alfonso Lozano. "Saliency of breast lesions in breast cancer detection using artificial intelligence". In: *Scientific Reports* 13 (2023). URL: https://api.semanticscholar.org/CorpusID:265403963.

Appendix A

A Foreign object masks

In order to obtain binary masks for foreign objects that differ from clips, a variety of approaches were employed, the details of which will be elucidated in the following sections.

To obtain a binary mask for pacemakers or other structures such as implants, it is first necessary to check whether there is a significant presence of bright pixels. This is achieved by computing the histogram to verify if at least 600 pixel values have the highest intensity value. This value was selected on the assumption that these objects should have larger areas. If this was the case, then the next step was to proceed with removing the object. Given that these objects have high-intensity values, a binary threshold with a value of 230 was applied. Subsequently, a dilation was performed with a kernel size of 11×11, followed by a closing operation with a kernel size of 9×9. The next step involved identifying the largest contours within the image. The resulting contour was then used to create the final binary mask, which had the same dimensions as the previous output image.

On the other hand, medium-sized foreign objects, such as wires, were removed after big foreign objects. The initial step was to directly apply a top hat filter to the image. The subsequent step involved applying a dilation filter with a kernel size of 7×7 for four iterations, followed by a closing filter with a kernel size of 9×9. A mean shift filter was then applied with a spatial and range radius value of 15 and 30, respectively. This was followed by a binary threshold filter with a value of 110, after which the largest contour in the image was obtained. The final step was to create a binary mask of the resulting contour.

Note that, these were two consecutive steps, they were applied in the same order as previously described.

B Image Segmentation Results

TABLE A.1
Image-Level DSC Statistics

Dataset	Model	Mean ± std	25%	50%	75%
Site A	U-NET	0.846 ± 0.121	0.841	0.873	0.905
	IP	0.613 ± 0.226	0.495	0.667	0.783
Site B	U-NET	0.684 ± 0.310	0.667	0.830	0.891
	IP	0.434 ± 0.239	0.229	0.477	0.629

Dice Similarity Coefficient (DSC) statistics at the image levels for both sites with clips. The statistics represent the median and standard deviation (std) and the 25%, 50% and 75% percentiles of the DSC distributions. Image processing algorithm (IP).

TABLE A.2
Region-Level DSC Statistics

Dataset	Model	Mean ± std	25%	50%	75%
Site A	U-NET	0.830 ± 0.194	0.833	0.880	0.920
	IP	0.567 ± 0.347	0.076	0.721	0.817
Site B	U-NET	0.718 ± 0.333	0.760	0.875	0.914
	IP	0.428 ± 0.320	0.000	0.558	0.693

Dice Similarity Coefficient (DSC) statistics at the region levels for both sites with clips. The statistics represent the median and standard deviation (STD) and the 25%, 50% and 75% percentiles of the DSC distributions. Image processing algorithm (IP).